(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117520552 B
(45)授权公告日 2024.04.16

(21)申请号 202410023000.7
(22)申请日 2024.01.08
(65)同一申请的已公布的文献号
申请公布号 CN 117520552 A
(43)申请公布日 2024.02.06
(73)专利权人 北京中科江南信息技术股份有限公司
地址 100086 北京市海淀区万泉河路68号8号楼1710室
(72)发明人 王巍 雷瑞恒 罗攀峰 曾纪才 韩魏 李宏超
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
专利代理师 郭美娇
(51)Int.Cl.
G06F 16/35(2019.01)
G06F 16/332(2019.01)
G06F 40/30(2020.01)
G06F 18/2415(2023.01)
G06F 40/284(2020.01)
(56)对比文件
CN 109493265 A,2019.03.19
CN 109635082 A,2019.04.16
CN 116562265 A,2023.08.08
KR 20230120711 A,2023.08.17
审查员 周亚楠

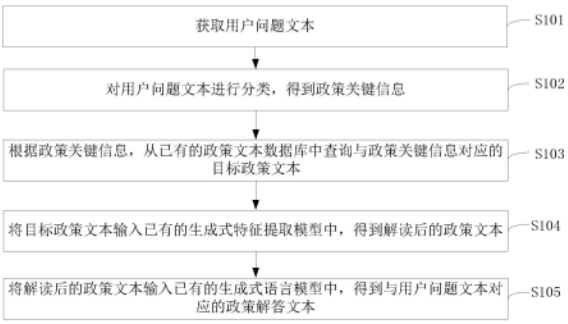
权利要求书2页 说明书13页 附图2页

(54)发明名称

政策文本处理方法、装置、设备及存储介质

(57)摘要

本公开提供一种政策文本处理方法、装置、设备及存储介质,涉及人工智能技术领域。在本公开的一些实施例中,获取用户问题文本;对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本;本公开结合人工智能技术,自动从政策文本数据库中获取到目标政策文本,利用生成式特征提取模型和生成式语言模型,自动得到用户问题文本的政策解答文本,降低人工成本,节省时间。



CN 117520552 B

1.一种政策文本处理方法,其特征在于,包括:

获取用户问题文本;

对所述用户问题文本进行分类,得到政策关键信息;

根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本;

将所述目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本,其中,所述生成式特征提取模型包括自注意力机制层,所述自注意力机制层包括第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,所述第一矩阵乘法层对所述目标政策文本进行处理,以得到第一特征,所述缩放层对所述第一特征进行处理,以得到第二特征,所述掩码层对所述第二特征进行处理,以得到第三特征,所述归一化层对所述第三特征进行处理,以得到第四特征,所述第二矩阵乘法层对所述第四特征进行处理,以得到所述解读后的政策文本;

将所述解读后的政策文本输入已有的生成式语言模型中,得到与所述用户问题文本对应的政策解答文本。

2.根据权利要求1所述的方法,其特征在于,所述对所述用户问题文本进行分类,得到政策关键信息,包括:

将所述用户问题文本输入BERT模型中,得到所述用户问题文本的语义信息;

将所述用户问题文本和所述语义信息输入贝叶斯网络上进行概率推理,得到贝叶斯网络中的多个节点的条件概率;

根据多个所述节点的条件概率,从多个所述节点中选择出条件概率满足设定概率条件的目标节点;

确定所述目标节点对应的政策关键信息。

3.根据权利要求1所述的方法,其特征在于,所述政策文本数据库为分层索引结构,所述根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本,包括:

根据所述政策关键信息,按照分层索引结构查询与所述政策关键信息对应的目标政策文本;所述分层索引结构对应的数据层,包括:热门数据层、历史数据层和稀有数据层;

其中,所述热门数据层,是第一查询等级的索引,包含第一查询频率的文档或条目的索引信息;

所述历史数据层,是第二查询等级的索引,包含所有政策文本的索引,在所述热门数据层中未查询到目标数据时,继续在所述历史数据层中查找所述目标数据。

所述稀有数据层,是第三查询等级的索引,包含第二查询频率的文档或条目的索引信息;

其中,所述第一查询等级查询优先级高于所述第二查询等级,所述第二查询等级的查询优先级高于所述第三查询等级;所述第一查询频率大于所述第二查询频率。

4.根据权利要求1所述的方法,其特征在于,所述政策文本数据库的获取步骤,包括:

获取原始政策文本;

对每个所述原始政策文本进行分词处理,得到每个所述原始政策文本对应的多个分词;

将每个所述原始政策文本对应的所述多个所述分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。

5.根据权利要求4所述的方法,其特征在于,所述深度学习模型包括:词性标注模型和命名实体识别模型,所述将每个所述原始政策文本对应的所述多个所述分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库,包括:

将每个所述原始政策文本对应的所述多个所述分词输入所述词性标注模型中,得到多个所述分词的词性;

将多个所述分词的词性输入命名实体识别模型中,得到多个所述分词的实体类型,齐其中,所述命名实体识别模型为双向长短期记忆网络,用于正序或者倒序读取句子文本。

6.一种政策文本处理装置,其特征在于,包括:

获取模块,用于获取用户问题文本;

分类模块,用于对所述用户问题文本进行分类,得到政策关键信息;

查询模块,用于根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本,其中,生成式特征提取模型包括自注意力机制层,所述自注意力机制层包括第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,所述第一矩阵乘法层对所述目标政策文本进行处理,以得到第一特征,所述缩放层对所述第一特征进行处理,以得到第二特征,所述掩码层对所述第二特征进行处理,以得到第三特征,所述归一化层对所述第三特征进行处理,以得到第四特征,所述第二矩阵乘法层对所述第四特征进行处理,以得到解读后的政策文本;

解读模块,将所述目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;

生成模块,将所述解读后的政策文本输入已有的生成式语言模型中,得到与所述用户问题文本对应的政策解答文本。

7.一种电子设备,其特征在于,包括:

处理器;

用于存储所述处理器可执行指令的存储器;

其中,所述处理器被配置为执行所述指令,以实现如权利要求1-5中任一项所述的方法中的各步骤。

8.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1-5中任一项所述的方法中的各步骤。

9.一种计算机程序产品,包括计算机程序/指令,其特征在于,所述计算机程序/指令被处理器执行时实现权利要求1-5中任一项所述的方法中的各步骤。

政策文本处理方法、装置、设备及存储介质

技术领域

[0001] 本公开涉及人工智能技术领域,尤其涉及一种政策文本处理方法、装置、设备及存储介质。

背景技术

[0002] 在进入新世纪的信息化时代,全球信息技术的快速发展及普及使得大数据、云计算和人工智能等新型信息技术不断进入到各个行业和领域中。人工智能从应用类型可分为四种:感知式AI与分析式AI应用较成熟,决策式AI近年来发展迅速,生成式AI迎来突破。生成式AI, 即AIGC(Artificial Intelligence Governance and Compliance),AIGC可实现更大数量、更高质量、更低单位成本,未来将从辅助创作生成趋向高度自动化自主创造。此外,AIGC将赋能多领域,加速人机共生的建设,迎接更多机遇与挑战。尤其在管理和公共服务领域,这些新技术带来了诸多优点,包括但不限于:决策的科学化、精确化,以及公共服务的便捷化、高效化。

[0003] 目前,对于政策文本需要人工进行处理,以得到用户可读性较强的政策文本。人力成本较高,耗时较长。

发明内容

[0004] 本公开提供一种政策文本处理方法、装置、设备及存储介质,以至少解决现有人力成本较高,耗时较长的技术问题。

[0005] 本公开的技术方案如下:

[0006] 本公开实施例提供一种政策文本处理方法,包括:

[0007] 获取用户问题文本;

[0008] 对所述用户问题文本进行分类,得到政策关键信息;

[0009] 根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本;

[0010] 将所述目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;

[0011] 将所述解读后的政策文本输入已有的生成式语言模型中,得到与所述用户问题文本对应的政策解答文本。

[0012] 可选地,所述对所述用户问题文本进行分类,得到政策关键信息,包括:

[0013] 将所述用户问题文本输入BERT模型中,得到所述用户问题文本的语义信息;

[0014] 将所述用户问题文本和所述语义信息输入贝叶斯网络上进行概率推理,得到贝叶斯网络中的多个节点的条件概率;

[0015] 根据多个所述节点的条件概率,从多个所述节点中选择出条件概率满足设定概率条件的目标节点;

[0016] 确定所述目标节点对应的政策关键信息。

[0017] 可选地,所述政策文本数据库为分层索引结构,根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本,包括:

[0018] 根据所述政策关键信息,按照分层索引结构查询与所述政策关键信息对应的目标政策文本;所述分层索引结构对应的数据层,包括:热门数据层、历史数据层和稀有数据层;

[0019] 其中,所述热门数据层,是第一查询等级的索引,包含第一查询频率的文档或目的索引信息;

[0020] 所述历史数据层,是第二查询等级的索引,包含所有政策文本的索引,在所述热门数据层中未查询到目标数据时,继续在所述历史数据层中查找所述目标数据;

[0021] 所述稀有数据层,是第三查询等级的索引,包含第二查询频率的文档或条目的索引信息;

[0022] 其中,所述第一查询等级查询优先级高于所述第二查询等级,所述第二查询等级的查询优先级高于所述第三查询等级;所述第一查询频率大于所述第二查询频率。

[0023] 可选地,所述生成式特征提取模型包括:自注意力机制层,所述自注意力机制层包括:第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,所述将所述目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本,包括:

[0024] 将所述目标政策文本输入所述第一矩阵乘法层,得到第一特征;

[0025] 将所述第一特征输入所述缩放层,得到第二特征;

[0026] 将所述第二特征输入所述掩码层,得到第三特征;

[0027] 将所述第三特征输入所述归一化层,得到第四特征;

[0028] 将所述第四特征输入所述第二矩阵乘法层,得到解读后的政策文本。

[0029] 可选地,所述政策文本数据库的获取步骤,包括:

[0030] 获取原始政策文本;

[0031] 对每个所述原始政策文本进行分词处理,得到每个所述原始政策文本对应的多个分词;

[0032] 将每个所述原始政策文本对应的所述多个所述分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。

[0033] 可选地,所述深度学习模型包括:词性标注模型和命名实体识别模型,所述将每个所述原始政策文本对应的所述多个所述分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库,包括:

[0034] 将每个所述原始政策文本对应的所述多个所述分词输入所述词性标注模型中,得到多个所述分词的词性;

[0035] 将多个所述分词的词性输入命名实体识别模型中,得到多个所述分词的实体类型。其中,所述命名实体识别模型为双向长短期记忆网络,用于正序或者倒序读取句子文本。

[0036] 本公开实施例还提供一种政策文本处理装置,包括:

[0037] 获取模块,用于获取用户问题文本;

[0038] 分类模块,用于对所述用户问题文本进行分类,得到政策关键信息;

[0039] 查询模块,用于根据所述政策关键信息,从已有的政策文本数据库中查询与所述政策关键信息对应的目标政策文本;

[0040] 解读模块,将所述目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;

[0041] 生成模块,将所述解读后的政策文本输入已有的生成式语言模型中,得到与所述用户问题文本对应的政策解答文本。

[0042] 本公开实施例还提供一种电子设备,包括:

[0043] 处理器;

[0044] 用于存储所述处理器可执行指令的存储器;

[0045] 其中,所述处理器被配置为执行所述指令,以实现上述的方法中的各步骤。

[0046] 本公开实施例还提供一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现上述的方法中的各步骤。

[0047] 本公开实施例还提供一种计算机程序产品,包括计算机程序/指令,所述计算机程序/指令被处理器执行时实现上述的方法中的各步骤。

[0048] 本公开的实施例提供的技术方案至少带来以下有益效果:

[0049] 在本公开的一些实施例中,获取用户问题文本;对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本;本公开结合人工智能技术,自动从政策文本数据库中获取到目标政策文本,利用生成式特征提取模型和生成式语言模型,自动得到用户问题文本的政策解答文本,降低人工成本,节省时间。

[0050] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,并不能限制本公开。

附图说明

[0051] 此处的附图被并入说明书中并构成本说明书的一部分,展示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理,并不构成对本公开的不当限定。

[0052] 图1为本公开示例性实施例提供的一种政策文本处理方法的流程示意图;

[0053] 图2为本公开示例性实施例提供的一种基于BERT模型的文本分类任务网络结构的示意图;

[0054] 图3为本公开示例性实施例提供的一种政策文本处理装置的结构示意图;

[0055] 图4为本公开示例性实施例提供的一种电子设备的结构示意图。

具体实施方式

[0056] 为了使本领域普通人员更好地理解本公开的技术方案,下面将结合附图,对本公开实施例中的技术方案进行清楚、完整地描述。

[0057] 需要说明的是,本公开的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不是用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本公开的实施例能够以除了在这里图示或描述的那些以外的顺序实施。以下示例性实施例中所描述的实施方式并不代表与本公开相

一致的所有实施方式。相反,它们仅是与本公开的一些方面相一致的装置和方法的例子。

[0058] 需要说明的是,本公开所涉及的用户信息包括但不限于:用户设备信息和用户个人信息;本公开中的用户信息的收集、存储、使用、加工、传输、提供和公开等处理,均符合相关法律法规的规定,且不违背公序良俗。

[0059] 在进入新世纪的信息化时代,全球信息技术的快速发展及普及使得大数据、云计算和人工智能等新型信息技术不断进入到各个行业和领域中。人工智能从应用类型可分为四种:感知式AI与分析式AI应用较成熟,决策式AI近年来发展迅速,生成式AI迎来突破。生成式AI, 即AIGC(Artificial Intelligence Governance and Compliance),AIGC可实现更大数量、更高质量、更低单位成本,未来将从辅助创作生成趋向高度自动化自主创造。此外,AIGC将赋能多领域,加速人机共生的建设,迎接更多机遇与挑战。尤其在管理和公共服务领域,这些新技术带来了诸多优点,包括但不限于:决策的科学化、精确化,以及公共服务的便捷化、高效化。

[0060] AIGC的优点在于它通过集成了大数据、云计算和人工智能技术,可以对政策进行智能化处理和解读,从而提供精确、及时的政策指导。在大数据分析的支持下,AIGC可以快速整合和处理海量政策数据,为决策提供数据支持。同时,利用云计算技术,AIGC能够实现高效、稳定的政策服务,不受时间和地点限制,方便公众获取。再者,人工智能技术的应用使AIGC能够对政策进行智能解读,提供更加精准的政策建议。然而,现有的AIGC系统也存在一些主要问题和不足。尽管通过人工智能和大数据技术可以对政策文本进行解析和理解,但由于政策语言的复杂性和特殊性,解析结果的准确性和全面性往往受到影响。不仅如此,大模型训练需要海量且优质数据,AI对数据训练集的消耗量远大于人类数据生产的速度,现有的政策解读系统在处理政策文本时往往需要大量的人工干预,这不仅消耗了大量的人力和时间,也可能因为人为因素影响解读结果的客观性和科学性。因此,如何实现政策解读的自动化处理,减少人工干预,是一个亟待解决的问题。

[0061] 针对上述技术问题,在本公开的一些实施例中,获取用户问题文本;对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本;本公开结合人工智能技术,自动从政策文本数据库中获取到目标政策文本,利用生成式特征提取模型和生成式语言模型,自动得到用户问题文本的政策解答文本,降低人工成本,节省时间。

[0062] 以下结合附图,详细说明本公开各实施例提供的技术方案。

[0063] 图1为本公开示例性实施例提供的一种政策文本处理方法的流程示意图。如图1所示,该方法包括:

[0064] S101:获取用户问题文本;

[0065] S102:对用户问题文本进行分类,得到政策关键信息;

[0066] S103:根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;

[0067] S104:将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;

[0068] S105:将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本。

[0069] 在本实施例中,上述方法的执行主体可以为终端设备或者服务器。

[0070] 其中,终端设备(terminal device)包括但不限于移动台(MS,Mobile Station)、移动终端(Mobile Terminal)、移动电话(Mobile Telephone)、手机(handset)及便携设备(portable equipment)等,该终端设备可以经无线接入网(RAN,Radio Access Network)与一个或多个核心网进行通信,例如,终端设备可以是移动电话(或称为“蜂窝”电话)、具有无线通信功能的计算机等,终端设备还可以是带无线收发功能的电脑、虚拟现实(VirtualReality,VR)终端设备、AR终端设备、工业控制(industrial control)中的无线终端、无人驾驶(self driving)中的无线终端、远程医疗(remote medical)中的无线终端、智能电网(smart grid)中的无线终端、运输安全(transportation safety)中的无线终端、智慧城市(smart city)中的无线终端、智慧家庭(smart home)中的无线终端等等,且终端设备上安装的操作系统包括但不限于:IOS、Android、windows、linux、Mac OS等操作系统。在不同的网络中终端可以叫做不同的名称,例如:用户设备,移动台,用户单元,站台,蜂窝电话,个人数字助理,无线调制解调器,无线通信设备,手持设备,膝上型电脑,无绳电话,无线本地环路台、电视等。为描述方便,本实施例中简称为终端设备。

[0071] 在本实施例中,对服务器的实现形态不作限定。例如,服务器可以是常规服务器、云服务器、云主机、虚拟中心等服务器设备。其中,服务器的构成主要包括处理器、硬盘、内存、系统总线等,和通用的计算机架构类型。

[0072] 在本公开的一些实施例中,需要首先建立政策文本数据库。一种可实现的方式为,获取原始政策文本;对每个原始政策文本进行分词处理,得到每个原始政策文本对应的多个分词;将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。本公开通过多重算法建立精细化的政策文本数据库,从文本数据中提取和表示有意义的特征。文本特征提取和表示的质量和准确性对于后续生成式特征提取模型的结果至关重要。

[0073] 本公开的一个核心技术是它的原理是将文本数据转化为机器可理解的特征表示,以便模型能够对文本进行理解、分类和标记,为用户提供准确和全面的解读结果。本公开将采用深度学习模型进行预训练,深度学习模型,例如,BERT模型。之后将采用词性标注模型进行词性标注,词性例如:名词、动词。词性标注模型是基于统计学习的词性标注器,主要采用了条件随机场模型。这是一种用于标签预测的统计学习模型,特别适合处理序列数据。之后,为了更好解读政策条规,将使用命名实体识别模型从文本中识别并分类特定类型的实体,如人名、地名、机构名、日期等。

[0074] 在本公开上述实施例中,将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。一种可实现的方式为,将每个原始政策文本对应的多个分词输入词性标注模型中,得到多个分词的词性;将多个分词的词性输入命名实体识别模型中,得到多个分词的实体类型。其中,命名实体识别模型为双向长短期记忆网络,用于正序或者倒序读取句子文本。

[0075] 图2为本公开示例性实施例提供的一种基于BERT模型的文本分类任务网络结构的示意图。结合图2上述过程主要涉及以下步骤:

[0076] 1,数据收集和准备:收集大规模的文本数据作为预训练数据。这些文本可以从各种公开的政策发布渠道获取,确保数据覆盖了模型所需的领域和语境。

[0077] 2,文本分词:对收集的文本数据进行分词操作,将句子分割成单词或子词的序列。BERT模型是一种基于Transformer架构的深度学习模型,用于进行预训练和生成文本的上下文相关表示。它通过无监督的方式从大规模的文本语料库中学习语义和语境信息,可以用于多种自然语言处理任务。BERT采用了一种特殊的分词方法,称为WordPiece分词,它将单词进一步分割成子词单位,以增加模型的词汇覆盖率。

[0078] 3,构建输入序列:对文本s分词后的文本t_1、t_2…t_n序列进行处理,构建BERT模型的输入序列。这包括添加特殊的标记,如[CLS]作为序列的开头,[SEP]作为不同句子之间的分隔符,以及填充序列以使其具有相同的长度。其中,图2中的Softmax是一种归一化函数。

[0079] 4,词性标注: 使用词性标注模型为每个单词分配一个词性(如名词、动词、形容词等),帮助后续模型更好的理解文本的词性。

[0080] 5,命名实体识别: 使用双向长短期记忆网络从文本中识别出特定类型的实体,如人名、地名、机构名、日期等。双向长短期记忆网络由两个 LSTM 层组成,正向 LSTM 层从左到右读取句子,而反向 LSTM 层从右到左读取句子。本公开采用双向长短期记忆网络能获取每个词其前面和后面词的信息,可以更好地理解句子的语义和语境,从而更准确地进行热点标记。

[0081] 6,模型训练:[CLS]对应的向量被输入到一个分类器。将准备好的输入序列和任务标签输入到BERT模型中进行训练。在训练过程中,通过最小化MLM任务和下游任务的损失函数来调整模型的参数。训练可以使用随机梯度下降等优化算法进行。

[0082] 7,微调和应用:预训练完成后,可以对BERT模型进行微调,以适应特定的下游任务。微调涉及在特定任务的训练数据上进行有监督的训练,调整模型参数以最大限度地提高任务性能。

[0083] 在本公开的一些实施例中,对用户问题文本进行分类,得到政策关键信息。一种可实现的方式为,将用户问题文本输入BERT模型中,得到用户问题文本的语义信息;将用户问题文本和语义信息输入贝叶斯网络上进行概率推理,得到贝叶斯网络中的多个节点的条件概率;根据多个节点的条件概率,从多个节点中选择出条件概率满足设定概率条件的目标节点;确定目标节点对应的政策关键信息。其中,本公开使用BERT模型将用户问题文本转换为一组高维特征向量,捕捉问题的语义信息,更好地理解用户的问题。之后使用随机森林分类,随机森林通过训练大量的决策树并通过投票的方式进行预测,可以有效地处理高维特征空间和过拟合问题。分类的结果包括政策主题、问题类型等。在识别和分类问题之后,将问题映射到已经构建的贝叶斯网络上。问题的各个方面对应于贝叶斯网络中的不同节点,基于问题映射到的节点和状态,使用贝叶斯网络进行概率推理,通过计算不同节点的条件概率,得到解答。需要说明的是,设定概率条件可以为概率最大。

[0084] 在生成式特征提取模型训练阶段,我们采用了双向长短期记忆网络进行特征抽取和语义理解。双向长短期记忆网络由两个LSTM层组成,正向LSTM从左到右读取句子,反向LSTM从右到左读取句子,这使我们能够更全面地捕捉政策文本中的语义和上下文关系。通过双向长短期记忆网络,能够更准确地标记和理解政策文本中的热点问题和关键信息,为后续的政策解读提供更有力的支持。通过预训练模型的隐藏层输出、注意力权重等方式来

实现。文本特征提取和表示的质量和准确性对于生成式特征提取模型的结果至关重要。它的原理是将文本数据转化为机器可理解的特征表示,以便模型能够对文本进行理解、分类和标记。在AIGC领域,常用的算法包括生成对抗网络、变分自编码器和循环神经网络等。每种算法都有其独特的优势和适用范围。生成对抗网络可以生成逼真的图片,变分自编码器可以生成多样性的图片,循环神经网络可以生成连贯的文本,本公开可以使用循环神经网络算法,具体步骤如下:

[0085] 1,分词和词向量化:首先,基于第一步已经将句子分割成子词,本公开将使用Word2Vec将分词后的单词转化为向量表示。它将单词映射到连续向量空间中,使得具有相似语义的单词在向量空间中距离较近。

[0086] 2,建立RNN模型: 对于一个句子或文本序列,需要将单词的向量表示组合成整个序列的表示。RNN模型特别适合处理序列数据,能够捕捉文本中的语义和上下文关系。本公开将选择长短期记忆(LSTM)结构并设置合适的参数。

[0087] 3,模型训练:选择合适、大量的政策文本,并使用相应的标签进行监督学习。本公开将使用梯度下降优化算法进行训练,通过调整参数使模型在训练集上的损失函数最小化。在训练过程中,还需要使用验证集来验证模型的泛化能力,并防止过拟合。

[0088] 4,推理分析:训练好的模型可以用来对新的政策文本进行推理分析。将新的政策文本输入模型,模型会输出相应的解读结果。之后可以根据模型的输出,提供精准的政策解读结果。

[0089] 5,反馈和持续优化:继续对模型进行优化,例如调整模型的参数,更新训练集,甚至改变模型结构。

[0090] 在本公开的一些实施例中,接收用户的政策解读请求,调用AIGC模块对相应的政策文本进行解析和解读,然后返回解读结果,可以实现对大量文本数据的高效处理和监测。

[0091] 在本公开的一些实施例中,根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本。一种可实现的方式为,根据政策关键信息,按照分层索引结构查询与政策关键信息对应的目标政策文本;分层索引结构对应的数据层,包括:热门数据层、历史数据层和稀有数据层;其中,热门数据层,是第一查询等级的索引,包含第一查询频率的文档或条目的索引信息;历史数据层,是第二查询等级的索引,包含所有政策文本的索引,在热门数据层中未查询到目标数据时,继续在历史数据层中查找目标数据。稀有数据层,是第三查询等级的索引,包含第二查询频率的文档或条目的索引信息;其中,第一查询等级查询优先级高于第二查询等级,第二查询等级的查询优先级高于第三查询等级;第一查询频率大于第二查询频率。

[0092] 传统信息库搜索一般使用倒排索引,倒排索引可能会占据计算机过大内存,以至于无法全部装入内存。如果查询都需要从磁盘上读取索引,就会导致查询效率降低。因此,本公开引入分层索引结构。其中,索引步骤如下:

[0093] 1,信息库检索:从政策文本数据库中筛选出相关的政策文本。本公开分层索引结构是将倒排索引分成两个或多个层级,每个层级都有自己的索引条目,用于提高查询效率和资源利用。举例如下,可以将分层索引结构对应的数据层分为三个层级:

[0094] 热门数据层: 这是最高层的索引,包含最频繁被查询的文档或条目的索引信息。这一层的索引包括了当前热门的政策文本或最常见的查询。在热门数据层中,索引条目的

更新频率较高,以确保反映最新的数据。这一层的目标是提供最快的响应时间。

[0095] 历史数据层:此层次包含更多的索引条目,覆盖了所有政策文本的索引。当查询在热门数据层中未找到时,系统会继续在历史数据层中查找。历史数据层的索引覆盖了全量的政策文本,但相对于热门数据层,响应速度可能略有降低。

[0096] 稀有数据层:如果系统需要进一步扩展索引容量,可以引入更深层的稀有数据层。这一层索引包含了不太常见的查询或更老的政策文本的索引条目。稀有数据层用于存储不经常访问的数据,以减少内存和计算资源的占用。这种分层索引结构允许系统平衡响应时间和资源占用。不同层级的索引可根据数据的重要性、查询频率和资源可用性进行管理和优化,以提供最佳的查询性能。这种结构有助于系统在处理大规模政策文本数据时,既保持高效的查询速度,又能够灵活应对不同类型的查询需求。之后对选定的政策文本导入BERT模型进行预处理,包括去除噪声数据、标准化文本、词语分词等。

[0097] 2,解读处理:利用Transformer模型对预处理后的政策文本进行解析和解读。它使用了自注意力机制,在处理语言的语义和上下文关系方面具有强大的能力。

[0098] 3,自动化处理:设计自动化的算法和系统,将预训练智能标签模型嵌入其中,以实现对大量文本数据的自动化处理。系统能够自动地对文本数据进行输入、处理和输出,节省时间和人力资源。

[0099] 在本公开的一些实施例中,生成式特征提取模型包括:自注意力机制层,自注意力机制层包括:第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本。一种可实现的方式为,将目标政策文本输入第一矩阵乘法层,得到第一特征;将第一特征输入缩放层,得到第二特征;将第二特征输入掩码层,得到第三特征;将第三特征输入归一化层,得到第四特征;将第四特征输入第二矩阵乘法层,得到解读后的政策文本。

[0100] 4,结果生成:接受政策文本数据并对其进行解析和解读。其输出是一个已解释和标记的政策文本,其中包括文本中的实体、关键信息、潜在问题等。

[0101] 5,反馈和持续优化:根据实际应用中的反馈信息,收集和分析模型的预测结果。根据反馈,进行系统和算法的持续优化和改进,以提高处理和监测的准确性和效率。

[0102] 在本公开的一些实施例中,将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本。例如,生成式语言模型可以为GPT-4模型。GPT-4模型被用于解读用户的问题和相关政策文本,以生成准确、连贯的回答。该模型基于大规模文本预训练,能够理解复杂的语义,并生成语言流畅、有深度的答案。

[0103] 本公开相比于现有技术,具有以下有益效果:

[0104] 一,精准解读:通过多重算法建立精细化的政策文本数据库,本公开可以准确解读政策文本中的语义和含义,为数据输入系统奠定了稳定的基础,可以为用户提供准确和全面的解读结果。

[0105] 二,自动化处理:相较于传统算法,本公开不仅使用倒排索引,还加入了自注意力机制,可以自动进行政策文本的解析和解读,无需人工干预,节省了人力成本和时间成本。

[0106] 三,快速响应:本公开在传统信息库搜索倒排索引中引入分层索引进行优化算法。这可以充分利用内存的高速度和磁盘的大容量,处理速度非常快,用户可以即时获得政策解读结果,提高了工作效率。

[0107] 在本公开上述方法实施例中,获取用户问题文本;对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本;本公开结合人工智能技术,自动从政策文本数据库中获取到目标政策文本,利用生成式特征提取模型和生成式语言模型,自动得到用户问题文本的政策解答文本,降低人工成本,节省时间。

[0108] 图3为本公开示例性实施例提供的一种政策文本处理装置30的结构示意图。如图3所示,该政策文本处理装置30包括:获取模块31,分类模块32,查询模块33,解读模块34和生成模块35。

[0109] 其中,获取模块31,用于获取用户问题文本;

[0110] 分类模块32,用于对用户问题文本进行分类,得到政策关键信息;

[0111] 查询模块33,用于根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;

[0112] 解读模块34,将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;

[0113] 生成模块35,将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本。

[0114] 可选地,分类模块32在对用户问题文本进行分类,得到政策关键信息时,用于:

[0115] 将用户问题文本输入BERT模型中,得到用户问题文本的语义信息;

[0116] 将用户问题文本和语义信息输入贝叶斯网络上进行概率推理,得到贝叶斯网络中的多个节点的条件概率;

[0117] 根据多个节点的条件概率,从多个节点中选择出条件概率满足设定概率条件的目标节点;

[0118] 确定目标节点对应的政策关键信息。

[0119] 可选地,政策文本数据库为分层索引结构,查询模块33在根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本时,用于:

[0120] 根据政策关键信息,按照分层索引结构查询与政策关键信息对应的目标政策文本;分层索引结构对应的数据层,包括:热门数据层、历史数据层和稀有数据层;

[0121] 其中,热门数据层,是第一查询等级的索引,包含第一查询频率的文档或条目的索引信息;

[0122] 历史数据层,是第二查询等级的索引,包含所有政策文本的索引,在热门数据层中未查询到目标数据时,继续在历史数据层中查找目标数据。

[0123] 稀有数据层,是第三查询等级的索引,包含第二查询频率的文档或条目的索引信息;

[0124] 其中,第一查询等级查询优先级高于第二查询等级,第二查询等级的查询优先级高于第三查询等级;第一查询频率大于第二查询频率。

[0125] 可选地,生成式特征提取模型包括:自注意力机制层,自注意力机制层包括:第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,解读模块34在将目标政策文本

输入已有的生成式特征提取模型中,得到解读后的政策文本时,用于:

[0126] 将目标政策文本输入第一矩阵乘法层,得到第一特征;

[0127] 将第一特征输入缩放层,得到第二特征;

[0128] 将第二特征输入掩码层,得到第三特征;

[0129] 将第三特征输入归一化层,得到第四特征;

[0130] 将第四特征输入第二矩阵乘法层,得到解读后的政策文本。

[0131] 可选地,政策文本数据库的获取步骤,包括:

[0132] 获取原始政策文本;

[0133] 对每个原始政策文本进行分词处理,得到每个原始政策文本对应的多个分词;

[0134] 将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。

[0135] 可选地,深度学习模型包括:词性标注模型和命名实体识别模型,将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库,包括:

[0136] 将每个原始政策文本对应的多个分词输入词性标注模型中,得到多个分词的词性;

[0137] 将多个分词的词性输入命名实体识别模型中,得到多个分词的实体类型。其中,命名实体识别模型为双向长短期记忆网络,用于正序或者倒序读取句子文本。

[0138] 关于上述实施例中的装置,其中各个模块执行操作的具体方式已经在有关该方法的实施例中进行了详细描述,此处将不做详细阐述说明。

[0139] 图4为本公开示例性实施例提供的一种电子设备的结构示意图。如图4所示,电子设备包括:存储器41和处理器42。另外,电子设备还包括电源组件43和通信组件44。

[0140] 存储器41,用于存储计算机程序,并可被配置为存储其它各种数据以支持在电子设备上的操作。这些数据的示例包括用于在电子设备上操作的任何应用程序或方法的指令。

[0141] 存储器41,可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。

[0142] 通信组件44,用于与其他设备进行数据传输。

[0143] 处理器42,可执行存储器41中存储的计算机指令,以用于:获取用户问题文本;

[0144] 对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本。

[0145] 可选地,处理器42在对用户问题文本进行分类,得到政策关键信息时,用于:

[0146] 将用户问题文本输入BERT模型中,得到用户问题文本的语义信息;

[0147] 将用户问题文本和语义信息输入贝叶斯网络上进行概率推理,得到贝叶斯网络中的多个节点的条件概率;

[0148] 根据多个节点的条件概率,从多个节点中选择出条件概率满足设定概率条件的目标节点;

[0149] 确定目标节点对应的政策关键信息。

[0150] 可选地,政策文本数据库为分层索引结构,处理器42在根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本时,用于:

[0151] 根据政策关键信息,按照分层索引结构查询与政策关键信息对应的目标政策文本;分层索引结构对应的数据层,包括:热门数据层、历史数据层和稀有数据层;

[0152] 其中,热门数据层,是第一查询等级的索引,包含第一查询频率的文档或条目的索引信息;

[0153] 历史数据层,是第二查询等级的索引,包含所有政策文本的索引,在热门数据层中未查询到目标数据时,继续在历史数据层中查找目标数据。

[0154] 稀有数据层,是第三查询等级的索引,包含第二查询频率的文档或条目的索引信息;

[0155] 其中,第一查询等级查询优先级高于第二查询等级,第二查询等级的查询优先级高于第三查询等级;第一查询频率大于第二查询频率。

[0156] 可选地,生成式特征提取模型包括:自注意力机制层,自注意力机制层包括:第一矩阵乘法层、缩放层、掩码层、归一化层和第二矩阵乘法层,处理器42在将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本时,用于:

[0157] 将目标政策文本输入第一矩阵乘法层,得到第一特征;

[0158] 将第一特征输入缩放层,得到第二特征;

[0159] 将第二特征输入掩码层,得到第三特征;

[0160] 将第三特征输入归一化层,得到第四特征;

[0161] 将第四特征输入第二矩阵乘法层,得到解读后的政策文本。

[0162] 可选地,政策文本数据库的获取步骤,包括:

[0163] 获取原始政策文本;

[0164] 对每个原始政策文本进行分词处理,得到每个原始政策文本对应的多个分词;

[0165] 将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库。

[0166] 可选地,深度学习模型包括:词性标注模型和命名实体识别模型,处理器42在将每个原始政策文本对应的多个分词输入深度学习模型中进行文本处理,得到文本处理后的政策文本数据库时,用于:

[0167] 将每个原始政策文本对应的多个分词输入词性标注模型中,得到多个分词的词性;

[0168] 将多个分词的词性输入命名实体识别模型中,得到多个分词的实体类型。其中,命名实体识别模型为双向长短期记忆网络,用于正序或者倒序读取句子文本。

[0169] 相应地,本公开实施例还提供一种存储有计算机程序的计算机可读存储介质。当计算机可读存储介质存储计算机程序,且计算机程序被一个或多个处理器执行时,致使一个或多个处理器执行图1方法实施例中的各步骤。

[0170] 相应地,本公开实施例还提供一种计算机程序产品,计算机程序产品包括计算机

程序/指令,计算机程序/指令被处理器执行图1的方法实施例中的各步骤。

[0171] 上述图4中的通信组件被配置为便于通信组件所在设备和其他设备之间有线或无线方式的通信。通信组件所在设备可以接入基于通信标准的无线网络,如WiFi,2G、3G、4G/LTE、4G等移动通信网络,或它们的组合。在一个示例性实施例中,通信组件经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,通信组件还包括近场通信(NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(RFID)技术,红外数据协会(IrDA)技术,超宽带(UWB)技术,蓝牙(BT)技术和其他技术来实现。

[0172] 上述图4中的电源组件,为电源组件所在设备的各种组件提供电力。电源组件可以包括电源管理系统,一个或多个电源,及其他与为电源组件所在设备生成、管理和分配电力相关联的组件。

[0173] 上述电子设备还包括显示屏和音频组件。

[0174] 显示屏包括屏幕,其屏幕可以包括液晶显示屏(LCD)和触摸面板(TP)。如果屏幕包括触摸面板,屏幕可以被实现为触摸屏,以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。触摸传感器可以不仅感测触摸或滑动动作的边界,而且还检测与触摸或滑动操作相关的持续时间和压力。

[0175] 音频组件,可被配置为输出和/或输入音频信号。例如,音频组件包括一个麦克风(MIC),当音频组件所在设备处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器或经由通信组件发送。在一些实施例中,音频组件还包括一个扬声器,用于输出音频信号。

[0176] 在本公开上述装置、电子设备和计算机程序产品实施例中,获取用户问题文本;对用户问题文本进行分类,得到政策关键信息;根据政策关键信息,从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本;将目标政策文本输入已有的生成式特征提取模型中,得到解读后的政策文本;将解读后的政策文本输入已有的生成式语言模型中,得到与用户问题文本对应的政策解答文本;本公开结合人工智能技术,自动从政策文本数据库中获取到目标政策文本,利用生成式特征提取模型和生成式语言模型,自动得到用户问题文本的政策解答文本,降低人工成本,节省时间。

[0177] 本领域内的技术人员应明白,本公开的实施例可提供为方法、系统、或计算机程序产品。因此,本公开可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本公开可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0178] 本公开是参照根据本公开实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0179] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指

令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0180] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0181] 在一个典型的配置中,计算设备包括一个或多个处理器 (CPU)、输入/输出接口、网络接口和内存。

[0182] 内存可能包括计算机可读介质中的非永久性存储器,随机存取存储器 (RAM) 和/或非易失性内存等形式,如只读存储器 (ROM) 或闪存(flash RAM)。内存是计算机可读介质的示例。

[0183] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括,但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带,磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质,可用于存储可以被计算设备访问的信息。按照本文中的界定,计算机可读介质不包括暂存电脑可读媒体 (transitory media),如调制的数据信号和载波。

[0184] 需要说明的是,在本文中,诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0185] 以上仅是本公开的具体实施方式,使本领域技术人员能够理解或实现本公开。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本公开的精神或范围的情况下,在其它实施例中实现。因此,本公开将不会被限制于本文的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

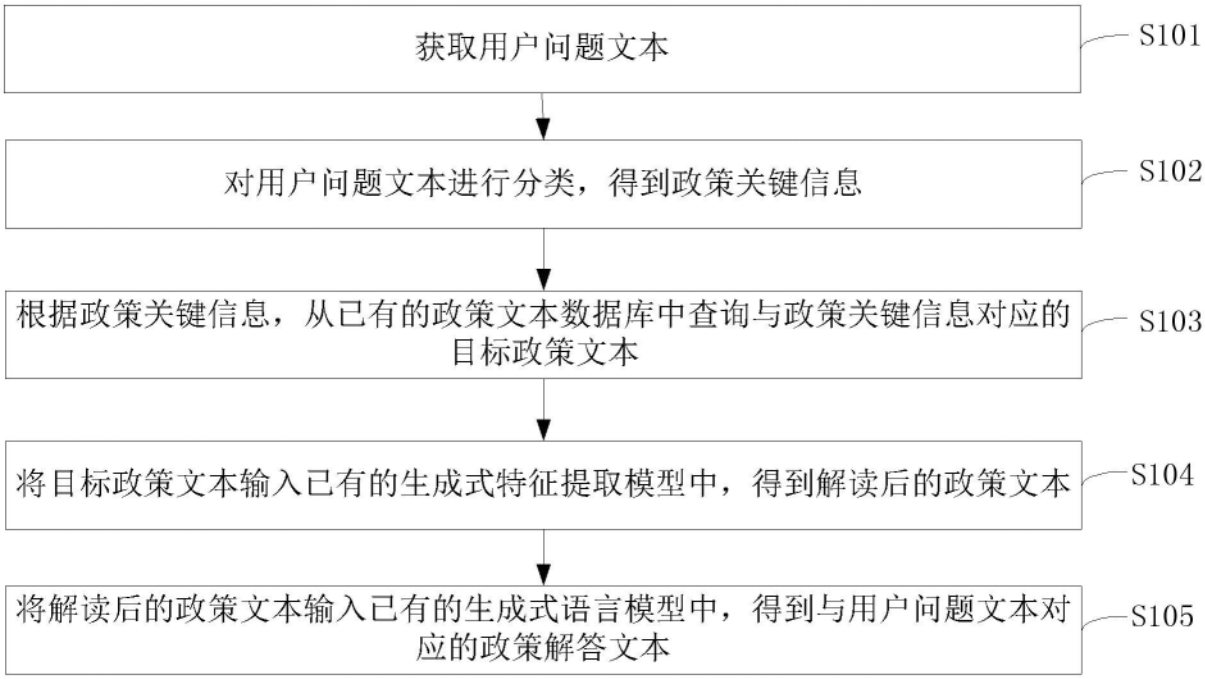

获取用户问题文本
S101
对用户问题文本进行分类，得到政策关键信息
S102
根据政策关键信息，从已有的政策文本数据库中查询与政策关键信息对应的目标政策文本
S103
将目标政策文本输入已有的生成式特征提取模型中，得到解读后的政策文本
S104
将解读后的政策文本输入已有的生成式语言模型中，得到与用户问题文本对应的政策解答文本
S105

图1

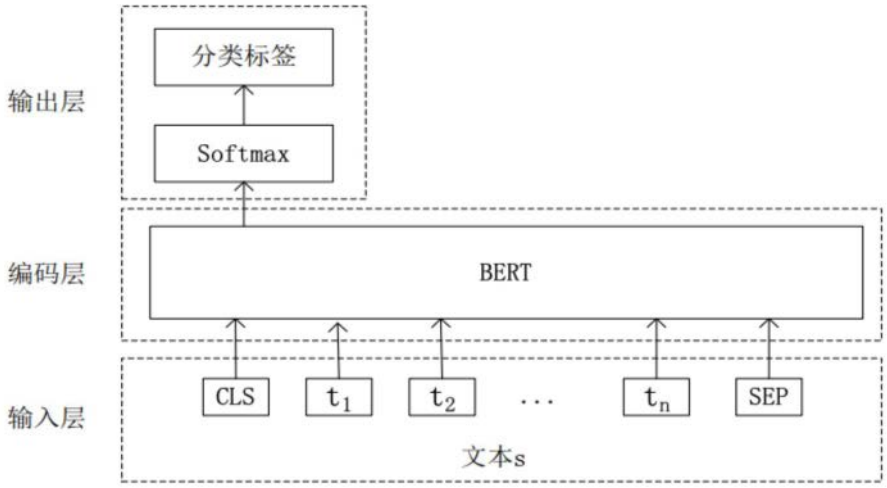

输出层
分类标签
Softmax
编码层
BERT
输入层
CLS
t_1
t_2
...
t_n
SEP
文本s

图2

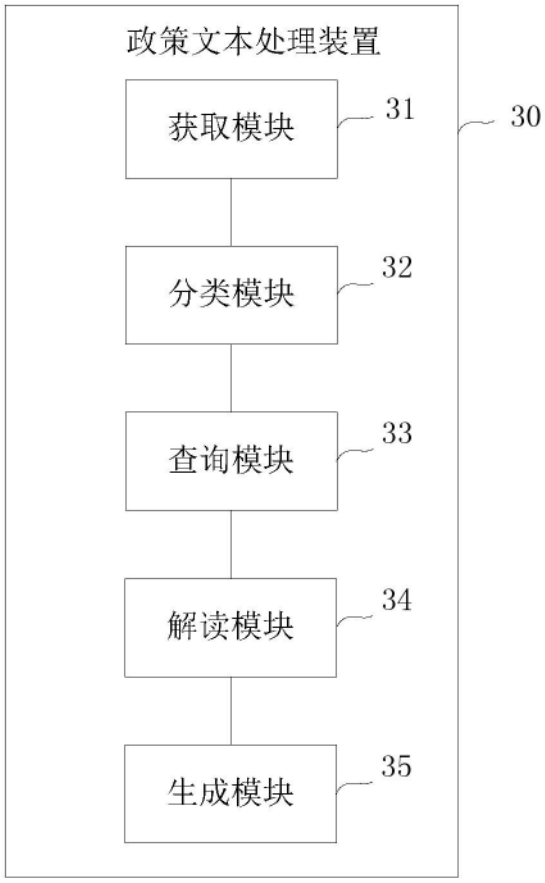

政策文本处理装置
获取模块
31
30
分类模块
32
查询模块
33
解读模块
34
生成模块
35

图3

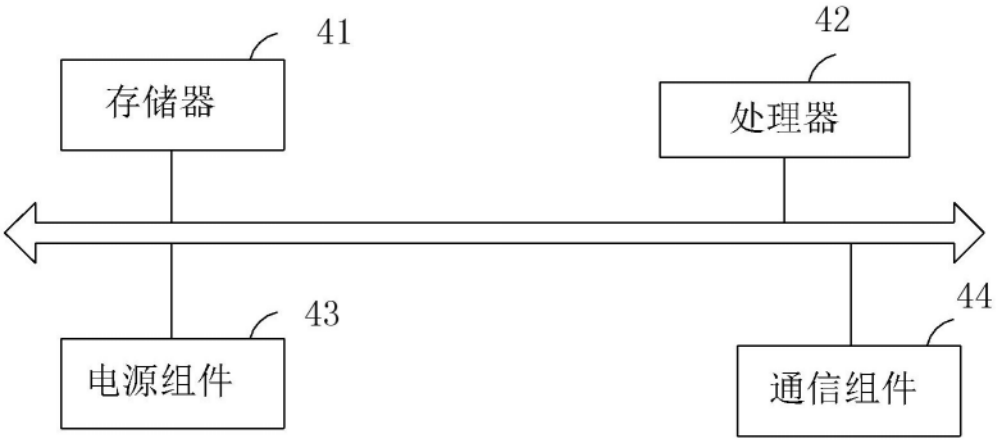

41
存储器
42
处理器
43
电源组件
44
通信组件

图4